CALIBRATION OF THE SYSTEM EVALUATION AND ESTIMATION OF RESOURCES SOFTWARE ESTIMATION MODEL (SEER-SEM) FOR THE AIR FORCE SPACE AND MISSILE SYSTEMS CENTER (SMC)

Thesis

Kolin D. Rathmann, Captain, USAF

AFIT/GCA/LAS/95S-9

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GCA/LAS/95S-9

CALIBRATION OF THE SYSTEM EVALUATION AND ESTIMATION OF RESOURCES SOFTWARE ESTIMATION MODEL (SEER-SEM) FOR THE AIR FORCE SPACE AND MISSILE SYSTEMS CENTER (SMC)

Thesis

Kolin D. Rathmann, Captain, USAF

AFIT/GCA/LAS/95S-9

Approved for public release; distribution unlimited

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense, the U. S. Government or the model developer.

AFIT/GCA/LAS/95S-9

CALIBRATION OF THE SYSTEM EVALUATION AND ESTIMATION OF RESOURCES SOFTWARE ESTIMATION MODEL (SEER-SEM) FOR THE AIR FORCE SPACE AND MISSILE SYSTEMS CENTER (SMC)

THESIS

Presented to the Faculty of the School of Logistics and

Acquisition Management

Air Education and Training Command

In Partial Fulfillment of the

Requirements for Degree of

Master of Science in Cost Analysis

Kolin D. Rathmann, B.S

Captain, USAF

September 1995

Approved for public release; distribution unlimited

Preface

This thesis effort was an evaluation of the accuracy of the System Evaluation and Estimation of Resources Software Estimation Model (SEER-SEM) for estimation of software effort. A calibration and a validation of the model was performed. It was assumed that the model would be an accurate predictor of actual software development effort. This assumption proved false. The model did not perform well due to the diversity of the data that was used. If an effort such as this is to be repeated, then the data used should be complete and be representative of a specific organization, not many.

I appreciate the support and assistance provided by my thesis advisor, Mr. Dan Ferens. He helped me keep this effort on target. I also want to thank Ms Sherry Stukes (MCR) for providing the actual database and for her help in explaining what the data meant, it made my task all the easier.

I would also like to thank my wife and children for both the support and the "play time" distractions. I could not have done it without you, nor would it have been any fun.

Kolin D. Rathmann

Table of Contents

List of Figures

Figure page

List of Tables

Abstract

This study's purpose was to determine whether calibration of the SEER-SEM impacted the effort estimates generated by the model for software developments. A historical database was provided by the Space and Missile Systems Center, Los Angeles, and used as the model's input data. The data was stratified into four usable platforms, military ground command/control, military ground signal processing, military specification avionics, and military mobile. Each platform's data sets were split, the majority of points for calibration of the model, and the rest for model validation.

The accuracy of SEER to this particular data set is limited, yet the model did respond to calibration. It is recommended that further calibration attempts be done within specific organizations. The diversity of the SWDB created too many factors for SEER to overcome.

CALIBRATION OF THE SYSTEM EVALUATION

AND ESTIMATION OF RESOURCES

SOFTWARE ESTIMATION MODEL (SEER-SEM)

FOR THE AIR FORCE SPACE AND MISSILE SYSTEMS CENTER (SMC)

I. Introduction

Overview

Software cost estimation is more an art than a science. Even with the expansion in software requirements, software estimation is still a comparatively primitive subject (Symons, 1991: xii).

As shown in Figure 1-1, if the present software cost growth rates of approximately twelve percent per year were to continue, then software costs in 1995 will be $36 billion for the Department of Defense (DoD), $225 billion for the US, and $450 billion worldwide (Boehm, 1987: 93).

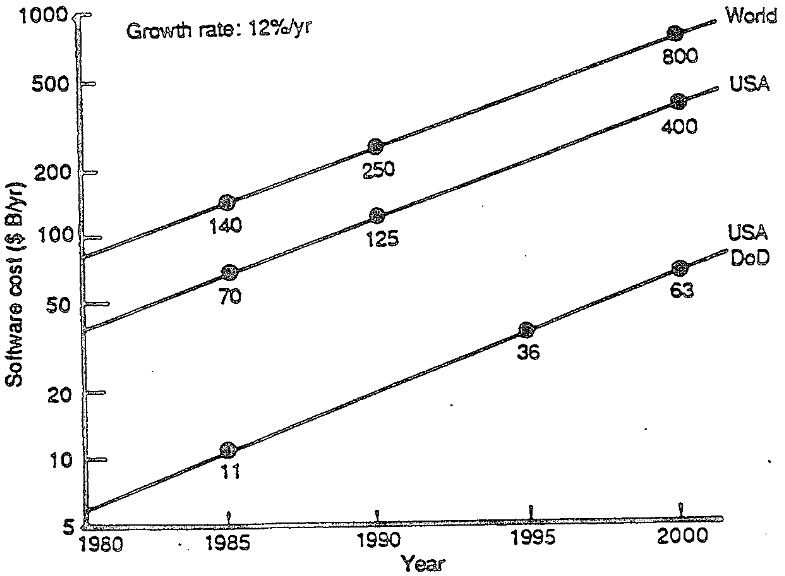


Figure 1-1 Software Cost Trends

Although conservative (actual DoD software costs in 1990 were around $34 billion), the previous figure gives the reader an appreciation for the rapid growth in software development and acquisition costs (Marsh, 1990: 62). The difference between Dr. Boehm's original estimate and the actual 1990 figure further strengthens the argument that software cost growth is spiraling "out of control." Software cost growth is not limited to the DoD. Software costs have become so high in the information systems industry that they, rather than technology or business needs, are driving the industry (Latamore, 1992: 100). The problem for both the DoD and industry is that there are good, useful technologies out there that will never be used because of costs (Latamore, 1992: 100). As our computer hardware capabilities have increased, it is our inability to produce reliable software at a reasonable cost that has been labeled by some as a "software crisis" which could ultimately impact our national security (Conte, 1986: 1; Symons, 1991: xi).

General Issue

With the declining defense budget, the DoD can no longer afford to continue business as usual. Software development costs are a substantial and growing portion of the defense budget (Christensen and Ferens, 1995: 1). If our nation is going to rely on smaller, yet technically sophisticated armed forces, we must better control software development and maintenance costs. Retired Air Force General Bernard Randolph, former Chief of Air Force Systems Command, characterized software as the "Achilles heel" of weapons development (Kitfield, 1989: 28). Representative John Murtha, D-PA, feels much the same way. To him software is the technology that threatens the nation's military dominance (Marsh, 1990: 62).

Having a highly technical mission, the Space and Missile Systems Center (SMC), Los Angeles Air Force Base, CA, is trying to control the software development cost estimating process in their system program offices (SPO) (Novak-Ley, 1994). According to Londeix, an estimate of software development costs is successful when:

1). The early estimate is within ± 30% of the actual final cost: this is the accuracy currently obtainable at an early stage of the development.

2). The method allows refinement of the estimate during the Software Life Cycle. A higher accuracy can be achieved by monitoring and re-estimating the development each time more information is available.

3). The method is easy to use for an estimator. This enables a quick re-estimate whenever it is necessary; for example during a progress meeting, the evaluation of alternatives in strategic choices.

4). The rules are understood by everybody concerned. Management feels more secure when the estimating procedures are easily understandable.

5). The method is supported by tools and documented. The availability of tools increase the effectiveness of the method, mainly because results can be obtained more quickly and in a standard fashion.

6). The estimating process can be trusted by software development teams and their management. This helps in gaining the participation of everybody concerned with the estimate. (Londeix, 1987: 3)

This effort will focus on tools and methods to control costs including the calibration and evaluation of the System Evaluation and Estimation of Resources Software Estimation Model (SEER-SEM).

Specific Issue

This research effort will evaluate the knowledge bases used in the SEER-SEM by calibrating them against SMC's database of over 2,000 diverse, historical software developments. The SEER-SEM cost estimating relationships (CERs) are consistent with the model developer's database, but are expected to be different when calibrated to SMC's database (Stukes, 1995). Calibration, in this effort, is adjusting the knowledge bases to the SMC database. Model accuracy is still not universally accepted. This research will determine SEER-SEM's ability to accurately predict software development effort when compared to actual data.

Research Objectives

This research addresses the following set of questions (Ourada, 1991: 1.4):

1. Given a credible set of actual DoD data, can the software cost-estimating model be calibrated? This question has already been addressed with regard to the SEER-SEM. The "Application Oriented Software Data Collection Software Model Calibration Report" proved that SEER-SEM can be calibrated (Apgar, Galorath, Maness, and Stukes, 1991: I-2).
2. Given a calibrated model, with another set of actual data from the same weapon system or other environments, can the model be validated?
3. Given a validated model, if another independent data set from another software development environment is used, is the estimate still accurate or more accurate than if not calibrated?

4. Is a calibration and validation of a model accurate for only specific areas of the weapon system application?

Scope of Research

Five different software cost estimating models are being calibrated for five thesis projects. Models covered include:

1. SEER-SEM
2. REVIC (Revised Enhanced Version of Intermediate COCOMO)
3. SLIM (Software LIfe Cycle Model)
4. PRICE-S (Programmed Review of Information for Costing and Evaluation Software)
5. SASET (Software Architecture, Sizing, and Estimating Tool)

Table 1-1 provides a summary of the model, developer and student responsible for each thesis effort.

Table 1-1 Software Cost Model Calibration Thesis Efforts

MODEL	DEVELOPER	STUDENT
SEER-SEM	GALORATH ASSOCIATES, INC.	Capt Kolin Rathmann
REVIC	RAY KILE	Ms Betty Weber
SLIM	QUANTITATIVE SOFTWARE MANAGEMENT (LARRY PUTNAM)	Capt Bob Kressin
PRICE-S	GE PRICE SYSTEMS	Capt James Galonsky
SASET	LOCKHEED MARTIN	1Lt Carl Vegas

Time constraints limited each thesis student to one model. This thesis will focus strictly on the SEER-SEM.

Summary

This chapter presented the scope of this research effort, breaking down the task from a general issue into specific research objectives. The state of software estimation in general will now be addressed in the literature review.

Definition of Terms

1. Calibration - The adjustment of selected parameters of the model to get an expected output with known inputs. In the world of statistics this effort is known as model building. For this research effort, the model already exists and will only be modified (Ourada, 1991: 1.6).

2. Validation - Testing a specific model using known inputs and establishing the output to within some error range. This is independent and non-iterative with calibration. This is often called cross-validation, in the world of statistics, since it will use a portion of an original data set kept out of the model (Ourada, 1991: 1.6).

3. Stratification - Breaking the SMC database into useable projects that contain the necessary information for the calibration effort. In this effort, stratification involved dividing the database along directed platform applications.

4. Source Lines of Code (SLOC) - do not include blank lines, comments, unmodified vendor supplied operating system or utility software, or other non-developed code. Include executable program instructions created by the project personnel which are delivered in the final product.

II. Literature Review:

Introduction

This section will cover the current state of software cost estimation, previous software cost model calibration efforts, and a description of the SEER-SEM.

Software Cost Estimation

There are many methodologies available for use in software cost estimation. Dr. Barry Boehm's paper, "Software Engineering Economics," provides a list of the major software cost estimation techniques:

1). Algorithmic Models - These methods provide one or more algorithms which produce a software cost estimate as a function of a number of variables which are considered to be the major cost drivers.

2). Expert Judgment - This method involves consulting one or more experts, perhaps with the aid of an expert-consensus mechanism such as the Delphi technique.

3). Analogy - This method involves reasoning by analogy with one or more completed projects to relate their actual costs to an estimate of the cost of a similar new project.

4). Parkinson - A Parkinson principle ("work expands to fill the available volume") is invoked to equate the cost estimate to the available resources.

5). Price-to-win - Here, the cost estimate is equated to the price believed necessary to win the job (or the schedule believed necessary to be first in the market with a new product, etc.).

6). Top-down - An overall cost estimate for the project is derived from global properties of the software product. The total cost is then split up among the various components.

7). Bottom-up - Each component of the software job is separately estimated, and the results aggregated to produce an estimate for the overall job. (Boehm, 1984: 7)

Table 2-1 provides an overall assessment of the strengths and weaknesses of the different software cost estimation methods.

Table 2-1
Strengths and Weaknesses of Software
Cost Estimation Techniques

Method	Strengths	Weaknesses
Algorithmic Model	◎ objective, repeatable, analyzable formula ◎ efficient, good for sensitivity analysis ◎ objectively calibrated to experience	◎ subjective inputs ◎ assessment of exceptional circumstances ◎ calibrated to past, not future
Expert Judgment	◎ assessment of representativeness, interactions, exceptional circumstances	◎ no better than participants ◎ biases, incomplete recall
Analogy	◎ based on representative experience	◎ representativeness of experience
Parkinson	◎ correlates with some experience	◎ reinforces poor practice
Price-to-win	◎ often gets the contract	◎ generally produces large overruns
Top-down	◎ system level focus ◎ efficient	◎ less detailed basis ◎ less stable
Bottom-up	◎ more detailed basis ◎ more stable ◎ fosters individual commitment	◎ may overlook system level costs ◎ requires more effort

(Boehm, 1984: 8)

The most important metrics for software cost estimation are source lines of code (SLOC) measurement and function points. SLOC measurement is a relatively simple way of measuring software volume. SLOC does not include blank lines, comments, machine generated or instantiated code, non-delivered test code, non-delivered debugging code, or begin statements from begin-end pairs (SEER-SEM User's Manual, 1994: 2-22). Source lines of code does include executable source lines such as all control, conditional, mathematical, declaration, input, and external output statements, as well as input/output formatting statements, deliverable job control, and debug and test code which is delivered in the final product (SEER-SEM User's Manual, 1994: 2-22). By concentrating on the executable statements in the source lines of code definition, the SLOC measurements are more likely to reflect the specification, design, coding and testing effort, and time scale of the software project (Londeix, 1987: 23). An estimating method based on software size expressed in SLOC presents numerous advantages according to Londeix:

1). the size is measurable with reasonable ease by using simple counting tools;

2). it is deliverable in the form of object code;

3). it is comparable across organizations on the basis of problem similarity without depending on the commenting capability of the individual programmers;

4). it does not depend on the design methodology;

5). it can be evaluated in a probabilistic fashion by a group of knowledgeable engineers. (Londeix, 1987: 26-27)

However, SLOC measurements have some rather serious drawbacks. SLOC lacks a standard definition for any major programming language, and there are more than 400 programming languages in use (Jones, 1994: 99). The software literature and even the lines of code counting standards are equally divided between those using physical lines and those using logical statements as the basis for the SLOC metric (Jones, 1994: 99). This metric is particularly dangerous, if used carelessly, because SLOC measurements penalize higher order languages, and the magnitude of the penalty is directly proportional to the level or power of the language (Jones, 1994: 99).

The function point metric is an alternative to the SLOC metric that is growing in use and popularity for software cost estimation. Function points originated with the work of Allan Albrecht as a methodology for estimating the size of a program by the number of functions the software was performing (Ferens and Gurner, 1994: 43). Based on his research, Albrecht further hypothesized that function points may be an alternative to using SLOC to estimate the cost or effort required for software development (Ferens and Gurner, 1994: 43-44). Function points are the weighted sum of five external attributes of software projects - inputs, outputs, inquiries, logical files, and interfaces - that have been adjusted for complexity (Jones, 1994: 100). The advantage of using the this metric is that function point total for an application does not change with the programming language (Jones, 1994: 100). Now it is possible to see the economic advantages of higher-order languages, such as Ada. Another advantage to the use of the function point metric is the continual improvement of function point theory and its practice by the International Function Point User's Group (IFPUG) (Ferens and Gurner, 1991: 49). Function points do have disadvantages. One problem is that Albrecht's five attributes are sometimes hard to define and count (Ferens and Gurner, 1991: 49). Another disadvantage of function points is that they are not readily adaptable to the real-time or scientific environments (Ferens and Gurner, 1991: 49).

The most prevalent issue discovered in the literature is related to the maturity level of software cost estimating. It terms of maturity, model estimation of software development cost and schedule may be only at level three, the defined stage (Ferens, 1994; handout). According to Humphrey, the defined process is still only qualitative: there is little data to indicate how much is accomplished or how effective the process is (Humphrey, 1989: 10).

In 1991, the Software Engineering Institute, a software think tank funded by the military, unveiled its Capability Maturity Model (CMM) which can grade the ability of a programming team to create predictably software that meets its customers' needs (Gibbs, 1994: 90). To date 261 organizations have been rated and according to the Software Engineering Institute the results have not been overwhelmingly positive:

> The vast majority - about seventy-five percent - are still stuck in level 1, they have no formal process, no measurements of what they do and no way of knowing when they are on the wrong track altogether. The remaining twenty-four percent of projects are at levels two or three. Only two elite groups earned the highest CMM rating, a level 5. (Gibbs, 1994: 90)

Besides maturity, much of the software cost inaccuracies can be attributed to the lack of cost estimating standardization. The lack of industry definitions for lines of code measures, cost per defect measures, and ratios established for programming sub-activities causes estimates to be inaccurate (Dreger, 1989: 4). According to Wellman, the lack of cost estimating standardization and establishment of good software development procedures and practices also contributes to software cost model error (Wellman, 1992: xvi). The software "crisis" is also attributable to poor software management particularly in the area of software estimation (Glass, 1994: 104) Glass cites three factors that contribute to software estimating deficiencies:

1. The software estimates are performed by many other management divisions but not the software people.
2. Software Estimates precede the requirements phase of the life cycle, that is,

before the software developer understands the problem to be solved.

3. Once obtained, estimates are held firm even when ongoing project experience tells us they are very bad and should be updated.

Software has suffered through most of its fifty-year history with inaccurate metrics and inadequate measurements (Jones, 1994: 100). Contributing to this problem is the apparent lack of interest from this nation's universities in producing qualified software engineers (Marsh, 1990: 63). Contrast software-intensive markets such as aircraft avionics, where the demand for software increases twenty-five percent each year, with an education system that graduates only four percent more software programmers annually (Kitfield, 1989: 28). Currently only twenty-eight universities offer graduate programs in software engineering; none offer undergraduate degrees (Gibbs, 1994: 95).

Software costs are spiraling out of control and somehow DoD must stop this trend (Kitfield, 1989: 30). This research will help DoD control escalating software development costs by providing a calibrated and validated tool: the SEER-SEM for SMC's software estimating requirements.

Previous Calibration Efforts

This research effort is a follow-on application of Ourada's thesis effort, the Coggins and Russell thesis effort, and a software model calibration report performed by Galorath Associates, Inc., and Management Consulting & Research, Inc. (Ourada, 1991; Coggins and Russell, 1993; Apgar, Galorath and Stukes: 1990). This effort will determine the accuracy of the model for estimating program costs using the newer version of the model.

The Coggins and Russell thesis effort did not specifically address the accuracy of the SEER-SEM model. The primary purpose of their effort was to document the differences in definitions, assumptions, and methodologies used by the REVIC, SASET, PRICE-S, and SEER-SEM cost models (Coggins & Russell, 1993: 4). They found that each cost model was unique in how it treated project phases, assumptions, schedules, and definitions.

The Ourada thesis is a comparison of various software cost models. However, in this case Ourada's purpose was the calibration, validation, and comparison of the different cost models (Ourada, 1991: 1.4). The four models he evaluated were REVIC, SASET, SEER-SEM, and COSTMODL. His findings indicated that the cost models where not very accurate for software estimation. In particular, he found the then current version of SEER-SEM to be uncalibratable and highly inaccurate against a historical database from SMC. This research effort will try to examine these findings.

The Apgar, Galorath, Maness and Stukes study was a calibration effort for three software models, PRICE-S, SEER-SEM, and SASET. The results for SEER-SEM indicated very good cost estimates, within 5.86% overall, once the model was calibrated (Apgar, Galorath and Stukes, 1991: III-1). However, the model developer performed the actual calibration. This research effort will independently derive calibration results for the SEER-SEM.

These previous efforts indicate that the use of software cost estimating models, particularly SEER-SEM, require more than simply inputting data and waiting for the results. An analyst, whose goal is an accurate prediction of a project's software development costs, must do more. Besides algorithmic cost models, several attempts are being made to estimate software costs using other estimating techniques or methods. Analog models are one of the other methods currently being used. This method is based on the comparison of a proposed software development project with one or more previous projects carried out in the same organization and for which the costs are known and understood (Wellman, 1992: 34-35). Another example of the techniques available to the software cost estimator is the use of prototyping models. Although not widely used, prototyping software developments has the potential to become a valuable aid in estimating particularly in circumstances where there is little data on which to base an estimate (Wellman, 1992: 40).

SEER-SEM Description

The model description is taken from the 1995 edition of the Space Systems Cost Analysis Group's (SSCAG) Software Methodology Handbook and the SEER-SEM User's Manual, version 4.0.

> SEER-SEM is part of a family of software and hardware cost, schedule and risk estimation tools. SEER models run on IBM, Macintosh, and Sun/UNIX platforms with no special hardware requirements. SEER-SEM is based on the mathematical software estimation model developed by Dr. Randall W. Jensen. SEER-SEM uses proprietary algorithms, some of which are found in the back of the User's Manual.
>
> SEER-SEM accepts SLOC or function points or both. When selecting function points, the user may use the International Function Point Users Group's (IFPUG) standard function points or SEER function based inputs which include

internal functions. Users follow a Work Breakdown Structure (WBS) describing each Computer Software Configuration Item (CSCI), Computer Software Component (CSC), and Computer Software Unit (CSU) (module or element) to be estimated. Knowledge bases are used to provide consistent inputs describing complexity, personnel capabilities and experience, development support environment, product development requirements, product reusability requirements, development environment complexity, target environment, schedule, staffing and probability. Users can modify all inputs to their specifications at any time.

The only parameter the user must enter is either an estimate of the functions or lines of code to give SEER-SEM a size figure on which to base its estimate. Software size is a primary driver for SEER-SEM.

In addition to cost analysis, SEER-SEM allows engineering alternative evaluation to select the optimum combination of schedule, cost, and personnel for the development project. SEER-SEM also allows for "what-if" analyses. By varying certain factors, such as required schedule, the analyst can quantify the impact to both cost and effort.

The model can also be calibrated to specific situations. Calibration of SEER-SEM involves the effort to customize input values to more closely reflect particular program characteristics. If historical data is available, SEER-SEM can be easily calibrated so that future projects will be fine tuned to that data SEER-SEM's Design to Technology and Design to Size functions provide the tools for calibration activities. (SSCAG Software Methodology Handbook; SEER-SEM User's Manual, 1994: 2-1-2-2)

Summary

This chapter discussed the state of software estimation in general. Some previous software cost estimating model calibration efforts were also discussed to provide a basis for this particular effort. Additionally this chapter included a description of the software estimating tool, SEER-SEM. Chapter 3 provides a detailed breakout of the assumptions and methodology used in calibrating SEER-SEM.

III. Methodology

Introduction

This chapter provides a description of the SMC database, addresses the steps required for model calibration and explains the statistical methods used to measure model calibration and validation accuracy. A contract for the database was established in 1989. The Space Division Comptroller (SD/ACC, now SMC/FMC, cost division of the Comproller's Office) contracted with Management Consulting & Research, Inc. (MCR) to consolidate and automate several existing government databases and to include other databases available from the Space Systems Cost Analysis Group (SSCAG), as well as many aerospace contractors and software developers (Apgar, Galorath and Stukes, 1991: I-1). In 1990, SD/ACC contracted with MCR to expand the depth and breadth of this database and to demonstrate its usefulness in the space software community. This consolidated, extended and expanded database is now known as the Space and Missile Systems Center Software Database (SWDB) (Fulton and Stukes, 1993: 2). The database is intended to be used for: analogy estimating; evaluation of new software architectures by comparing size and effort parameters with historical programs; development of Cost Estimating Relationships (CERs) and parametric models; and the calibration of existing software estimating models in use by SMC (Apgar et al, 1991: I-1, I-2).

Data

According to the SSCAG Software Methodology Handbook:

> The SMC SWDB currently contains records of information. These records contain extensive information about software development projects. Each record

> has up to 273 fields of information representing items such as: Operating Environment, Software Application, Software Function, Development Language. It includes Schedule, Sizing, Effort, Maintenance Information as well as detailed attributes defining the software.

The SMC database easily accommodates queries to search for specific types of software developments. Information in the database included size, effort, schedule and various development information. However, many of the records in the database lack some important information for several of the SEER-SEM variables (not an inherent limitation of the database but a lack of attention to detail by the contributors of the software data). SEER's default parameters for the various software development attributes were used in those situations where there was no other data available to improve the estimate. For a database record to be considered for calibration, it had to contain reasonable size and effort information. For example, a data point containing 50,000 lines of code developed in 6 person months was eliminated as unreasonable. Normalized size and effort were the parameters used in this research project. The normalization procedures and assumptions are documented in other sources (Stukes and Apgar, 1994: F-1-F-3). Some program data points were insufficient for calibration (i.e., no effort actuals), while other data points appeared to be outside the logical scope for the particular industry (i.e., extremely high or low productivity). The following criteria were used to eliminate inappropriate data records:

1. Actual Effort = 0
2. Actual Size = 0
3. Data provided at the project not the CSCI level

4. CSCI size less than 2,000 lines of code (Stukes,1994:V-15; McRitchie, 1995)

5. CSCI size greater than 150,000 lines of code (Stukes, 1994: V-15; McRitchie, 1995)

6. Areas of Operation (Platforms) with no statistical significance (i.e. After grouping the CSCIs by Platform, the Platforms that had less than eight data points were eliminated).

7. Platform or Application Area outside the scope of this thesis effort as stated by the sponsor of this research, SMC.

8. Software developments by foreign nations. (i.e. European Space Agency data)

Assumptions:

1. If a project did not specify the development standard used, then 2167A was assumed.

2. This effort assumed that there were one hundred fifty-two hours per person month. If a record stated some other basis for a person month, then it was normalized. If the hours per person month field in a record was left blank, it was assumed to have used one hundred fifty-two hours per person month.

3. If no development method was specified (i.e. incremental, spiral, waterfall, etc.) then no method was assumed. Instead the SEER "no knowledge" knowledge base was used for development method. This selection does not change any project software parameters.

4. The normalized effort results from the SWDB were assumed to correctly account for the labor categories and acquisition phases included in each activity.

5. Normalized effective size (in terms of SLOC) from the SWDB was assumed throughout this effort. The normalized effective size is equal (if both new and pre-existing size was reported) to reported pre-existing size times 40% of the re-design percentage plus 25% of the re-implementation percentage plus 35% of the re-tested percentage (Stukes and Apgar, 1994: F-2-F-3).

6. This effort assumed peak staffing profiles were valid and used them when provided.

7. During the data compilation, several projects listed "Modern Development Practices Use" as high to very high. This seemed very optimistic compared to the development team's capabilities (as stated in the respective record), and were kept at the knowledge base values.

8. This effort assumed that the projects contained in the database were optimized for schedule. Optimizing for schedule (min time) assumes the development will be finished as quickly as possible (SEER-SEM User's Manual, 1994: 8-33). SEER allows the analyst to choose between optimizing schedule and effort. SMC's recommendation was to optimize on schedule, since it more accurately represents DoD development efforts.

Methodology:

This research effort will use the SMC managed database containing roughly 2,000 different projects from the Air Force, Army, Navy, NASA, and others.(Stukes, 1995). The SEER-SEM can estimate based on either the analyst's detailed input from the various projects or from the model's default values, if software project size is provided (SEER-SEM User's Manual, 1994: 2-1). In SEER-SEM, there are default knowledge bases. SEER-SEM knowledge bases provide the inputs that are key to making estimates (SEER-SEM User's Manual, 1994: 2-6). There are six types of knowledge bases, which describe six different categorizations of the project WBS :

1). Platform - describes the primary operating environment of the project, such as avionics, business, ground based, manned space, missile, mobile, ship or unmanned space.

2). Application - describes the overall function of the software, such as computer-aided design (CAD), command and control, database, management information systems (MIS), office automation, radar, simulation, etc.

3). Acquisition Method - describes how the software project will be acquired. Is the project an "in-house", new work effort, is it a modification, re-engineering, purchase and integration, maintenance only, etc.?

4). Development Method - describes the development methods to be used during development. This knowledge base includes Ada, spiral, prototyping, object oriented design, evolving, traditional incremental or traditional waterfall.

5). Development Standard - describes the documentation, quality, and test standards to be followed. This knowledge base includes commercial and government standards.

6). Class - describes the type of class of software a WBS item belongs in. This category is convenient for creating user defined knowledge bases.

According to the SEER-SEM User's Manual, each knowledge base loads certain parameters which are appropriate to the individual knowledge area (SEER-SEM User's Manual, 1994: 7-1). The manual continues by describing the process in which knowledge base parameters are used by SEER-SEM,

> The Platform knowledge base is the first knowledge base loaded; it loads all of the parameters. The Application knowledge base loads next, and overwrites some parameters which were loaded by the Platform (knowledge base) with more specific information. The next knowledge base to load is the Development Method, and then the Development Standard, each overwriting parameters loaded by previous knowledge bases where applicable.
>
> Somewhat different are the Class knowledge bases. These are reserved for users to create custom knowledge bases. The Class knowledge base loads after Development Standard, and is the last knowledge base to be loaded, so its parameters take precedence over all other knowledge bases. This is where users can include their own labor rates, tools, practices, calibration and other information in this area. (SEER-SEM User's Manual, 1994: 7-1)

This research was conducted in two parts: model calibration and validation. The calibration began by stratifying the SWDB into the platforms specified by the sponsors of this thesis effort, SMC. The platforms selected include the following: unmanned space, military mobile, military ground, missile and military specification avionics. The military ground platform was further stratified by software application either command and control or signal processing again based on the needs of the sponsor as well as the availability of data.

Initially this effort was aimed at calibrating SEER's effort adjustment factor, schedule adjustment factor, and the effective technology rating by platform. After early analysis, based on the previously mentioned reasonableness criteria of the stratified platform data, two things became apparent. First, much of the data sorted by platform did not satisfy the requirements of useful information as defined above for both effort and schedule calibration. Most data points did not include schedule information so this factor will not be calibrated in this effort. Secondly, the diversity of the submitting organizations' missions, personnel attributes and performance made calibrating an effective technology rating impractical. An effective technology rating calibration is more practical within specific offices or programs not in cases where the calibration is attempting to cross Service, Governmental or Non-Governmental lines (McRitchie, 1995). As mentioned earlier, this database contains records from the Air Force, Army, Navy, NASA, other governmental agencies and private industry. The records used in this calibration effort display this same diversity.

The number of records used for calibration depended on the number of total, yet reasonable, data points available from each platform stratification. If the stratification resulted in between nine and eleven data points, then eight were used for calibration and the remainder for validation. If twelve of more data points were available, then two-thirds were used for calibration and the rest for validation. If a platform stratification resulted in fewer than eight data points, the platform was excluded from further analysis because of the limited data available. The missile and unmanned space platform stratifications were a victim of this restriction. The missile platform provided five data points of which only four were usable. The unmanned space platform only provided three data points. These platforms will not be calibrated or validated since any conclusions drawn from such limited data sets would be questionable.

The remaining platform data sets had their respective records sorted by size and then the data points for calibration were selected randomly. The randomization procedure was accomplished by starting with the smallest data point (in normalized effective size terms), selecting two points, then skipping one and repeating until all points in the platform data set where exhausted.

The records used for calibration were entered into the SEER-SEM using as much of the actual information contained in a particular record as possible. SEER's parameters were adjusted based on the actuals or they were left at their default values. The only two parameters adjusted, if not provided, were the development method knowledge base which was set to "no knowledge" and the development standard knowledge base which was set to 2167A.

The calibration technique used was the process suggested in the SEER-SEM User's manual, version 4.0. SEER's calibration/design-to mode was used to input actuals and then compared to SEER's estimates for effort. The resulting difference between the actual effort and the estimated effort impacts the calculated effort calibration adjustment factor. This factor has a default value of 1.0. This procedure is repeated for each record and then the average effort adjustment factor is used to create a custom knowledge base to use in the validation.

One area worth mentioning at this point is the schedule calibration feature in SEER-SEM. This effort did not use this calibration technique because in most of the projects schedule was not known. The differences between actuals and estimates affecting the effort calibration adjustment do not impact the schedule adjustment factor.

Throughout this exercise, the schedule adjustment factor stayed at its default value of 1.0. To impact the schedule adjustment factor, actual schedule information would need to be known and then it could be used similarly to the effort calibration in the SEER-SEM calibration/design-to mode.

During the validation, the remaining data points are used. The validation data points are put in the same way as the calibration data points. Any further information included in a particular record is also input into the SEER-SEM. Also the same assumptions were used for any missing data elements. The only difference in the validation methodology is the additional selection of the class knowledge base. The class knowledge base contains the custom calibration effort adjustment factors from the calibration mode. A calibration effort adjustment factor was created for each platform and for the command and control and signal processing software application environments in military ground.

The objective is to examine the statistical consistency when comparing the known output to the estimated output of the model. This validation should show that the model is accurate within 25%, 75% of the time for the effort in the environment of the calibration (Conte, 1986: 173)

To test the accuracy of the models, several statistical tests are used. The first tests are the magnitude of relative error and mean magnitude of relative error. The equation for magnitude of relative error (MRE) is Equation 3.1 and for mean magnitude of relative error (MMRE), Equation 3.2. A small value of MRE indicates SEER-SEM is predicting accurately. The key parameter however, is MMRE. For SEER-SEM to be considered acceptably accurate, MMRE should be less than or equal to 0.25 (Conte, 1986: 148-176). The use of MRE and MMRE relieve the concerns of positive and negative errors canceling each other and giving a false indication of model accuracy.

$$MRE = \left| \frac{E_{act} - E_{est}}{E_{act}} \right| \qquad \text{Eq. 3.1}$$

$$MMRE = 1/n * \sum_{I=1}^{n} MRE_i \qquad \text{Eq. 3.2}$$

Errors using the MRE and MMRE tests can be of two types: underestimates, where $E_{est} < E_{act}$; and overestimates, where $E_{est} > E_{act}$. Both errors can have serious impacts on estimate interpretation. Large underestimates can cause projects to be understaffed and, as deadlines approach, project managers will be tempted to add new staff members, resulting in a phenomenon known as Brook's law: "Adding manpower to a late software project makes it later". Large overestimates can also be costly, staff members become less productive (Parkinson's law: "Work expands to fill the time available for its completion") or add "gold-plating" that is not required by the user (Kemerer, 1987: 420)..

The second set of statistical tests are the root mean square error (RMS), Equation 3.3, and the relative root mean square error (RRMS), Equation 3.4. The smaller the value of RMS the better the model's ability to forecast actual performance. For RRMS, an acceptable model will give a value of RRMS < 0.25 (Conte, 1986: 175).

$$RMS = 1/2\ (1/n * \sum_{I=1}^{n} (E_{act} - E_{est})2) \qquad \text{Eq. 3.3}$$

$$RRMS = \frac{RMS}{1/n * \sum_{I=1}^{n} E_{act}} \qquad \text{Eq. 3.4}$$

The third statistical test used is the prediction level test, Equation 3.5, where k is the number of projects in a set of n projects whose MRE is less than or equal to ± 25 percent. If a project's MRE is less than or equal to ± 25 percent, then the project receives a "counter" value of 1. If a project's MRE does not fall between ± 25 percent, then it receives a "counter" value of 0. Then next step is to sum the "counter" values and divide by the number of projects.

$$PRED\ (X) = k/n \qquad \text{Eq. 3.5}$$

For example, if PRED (0.25) = .83, then 83% of the predicted values fall within 25% of their actual values. To establish the model accuracy, 75% of the predictions must fall within 25% of the actual values, or PRED (0.25) >= 0.75 (Conte, 1986: 173).

The fourth statistical test used was the Wilcoxon Signed-Rank Test. The Wilcoxon signed-rank test is a non-parametric test and was used to test for bias in the distributions of SEER's estimates of project effort for the validation data set. The validation data set was used both calibrated and in uncalibrated form then compared to the actual observations. This was accomplished by using Statistix Version 4.0 software package. This procedure tests the hypothesis that the frequency distributions for the two groups are identical. The absolute value of the differences were first ranked from the least to the greatest. If the data were truly unbiased, one would expect that just as many negative differences occur as positive differences, thereby the number of positive and negative differences would sum to zero. Differences that are near zero (absolute value less than 0.00001) are ignored and tied values are given a mean rank (Statistix User's Manual, 1992: 111). The Statistix User's Manual states that differences are considered to be tied if they are within 0.00001 of one another. "Sizable differences in the sums of the ranks assigned to the positive and negative differences would provide evidence to indicate a shift in location between the distributions (Mendenhall, et. al., 1990: 680). In effect, if there exists a significant difference in the sums of the ranks assigned to the positive and negative differences, one could conclude that the estimate observations, when compared to the actuals, are bias toward being either high or low. The signed rank test tests the null hypothesis that the median of the differences equals zero.

The exact p-values for the Wilcoxon signed rank test are computed by Statistix for small to moderate sample sizes (20 or fewer cases) (Statistix User's Manual, 1992: 112). Since the data sets used in this research were relatively small, n less than 20, Statistix was used to calculate the exact p-values. The exact p-value for a two sided test is computed by doubling the one sided p-value. The User's manual warns that when ties are found to be present in the data the "exact probability" is no longer exact but will usually be a good approximation. Dependent upon the results of the Statistix's calculated p-values, one could then conclude that SEER-SEM is biased.

Summary

This chapter reviewed the data that was used for this research effort and the technique to perform the calibration and validation of SEER-SEM. The statistical techniques used were also presented. Chapter IV is the presentation of the analyses and results of the methodologies and assumptions used in this research process.

IV. Analysis and Results

Introduction

This chapter will present an analysis and result of the research effort. An analysis of the database will be presented and then the platform calibration and validation results.

Data

This part of the research effort was surprising difficult. It was expected that with over 2,600 records in the SWDB, platform stratifications would yield sizable numbers of usable projects for both calibration and validation. This was not the case. Because of the limited quantity of good data (as described in Chapter 3), two platform stratifications, missile and unmanned space, did not generate sufficient data points for calibration and validation. The remaining stratifications provided fewer data points than expected but calibration and validation still took place.

Another problem occurred late in this research effort. The problem was that most of the unmanned space data points where incorrectly included in this category. The data points should have been included in a category called ground in support of space. The SWDB includes this category in its available platform applications. However, SEER-SEM does not have such a category so they were included in the military ground applications. Platform is a significant cost driver in SEER and any data points not aligned with their proper categories will likely impact the accuracy of the estimates that SEER-SEM generates.

Lastly, the records that were used often contained limited information. Records of this type included only platform, software application, size and effort. SEER allows the user to go beyond these basic inputs by including many additional project attributes that the user can modify. The suppliers of the data to the SWDB often times do not describe the particular attributes of their respective projects. They do not comment on the project's software attributes, personnel attributes or the software environment; all of which can significantly impact SEER's effort estimates. If the knowledge base settings are drastically different than the actual environment in which the project took place then the estimates can be grossly different from the reported actuals..

Calibration

The data points that make up the various platform stratifications were entered into the model to test for model accuracy for each respective environment. The estimates and the actuals were graphed against size to see if the data exhibited the expected positive growth in effort with increased project size. The data did exhibit this trend which is important since SEER-SEM relies heavily on effort to size for its estimate computations (McRichie, 1995). The graphs of effort to size are included in Appendix A. Table 4.1 shows the statistical results of this stage in the analysis. The statistical results for each record used in the model calibration are provided in Appendix B.

Table 4.1 SEER-SEM Calibration Accuracy Results

	Mil-Spec Avionics	Military Ground C/C	Military Ground Sig Processing	Military Mobile
MMRE	.9233	.5307	1.44	2.8022
RMS	552.8	253.7	326.1	771.5
RRMS	1.472	1.031	1.082	3.711
PRED (.25)	25.00%	31.25%	6.25%	11.11%

The model's ability to accurately estimate these particular data sets is limited. The best the model could do was to estimate within 25% only 31.25% of the time. The results support the need for calibration. For reference, Conte points out that a model's estimates are statistically acceptable when MMRE and RRMS are less than .25, RMS is small (approaching 0) and PRED (.25) is greater than or equal to 75% (Conte, 1986: 150-176).

If the estimates created by SEER-SEM differed from the actuals, then a calibration effort adjustment factor was calculated for each record of each respective platform. This calibration effort adjustment factor is simply a number that when multiplied by the SEER-SEM estimate equals the actual effort. For example, if the actual effort for a particular project was four person months and the SEER-SEM estimate was two person months, then the calibration effort adjustment factor would equal two. This process was repeated for each record making up the platform stratifications. Table 4.2 shows the mean calibration adjustment factor used for each platform. The calibration effort adjustment factors for each record used in the calibration are provided in Appendix C.

Table 4.2 Platform Mean Calibration Effort Adjustment Factors

	Mil-Spec Avionics	Military Ground C/C	Military Ground Sig Processing	Military Mobile
Calibration Effort Adjustment Factor	0.85	1.17	1.36	1.20

Validation

The validation process involved creating a custom knowledge base in order for the above calibration effort adjustment factors to used. A Class knowledge base was created for each platform. The only parameter adjusted or used was the calibration effort adjustment factor in this knowledge base. The validation data points were entered into the model in the same fashion as the calibration data set. Initially, the validation data points were run through SEER without using the Class knowledge base. This provided a baseline for later comparison. The comparison was important for determining if the effort adjustment factor calibration impacted the accuracy of the model. Next, the validation data set was updated with one additional knowledge base, the custom or Class knowledge base containing the calibration effort adjustment factor. Table 4.3 shows the statistical results of SEER's estimates without using the Class knowledge base, while Table 4.4 shows the results of SEER's estimates using the Class knowledge base. The actual results for each record are provided in Appendix B.

Table 4.3 SEER-SEM Validation Accuracy Results without Class Knowledge Base

	Mil-Spec Avionics	Military Ground C/C	Miltary Ground Sig Processing	Military Mobile
MMRE	.459	.314	1.5398	.3903
RMS	89.73	81.05	500.5	158.7
RRMS	.337	.259	1.278	.28
PRED (.25)	0.00%	42.86%	28.57%	25.00%

Table 4.4 SEER-SEM Validation Accuracy Results with Class Knowledge Base

	Mil-Spec Avionics	Military Ground C/C	Military Ground Sig Processing	Military Mobile
MMRE	.2403	.3108	2.092	.462
RMS	63.92	92.55	630.5	193.9
RRMS	.24	.296	1.61	.342
PRED (.25)	100.00%	28.57%	42.86%	25.00%

The validation process proved that SEER-SEM is calibrateable, however the results from the calibration do not show much improvement over the SEER's uncalibrated validation data set estimates in comparison to the calibrated data set. Although the Military Specification Avionics platform shows that SEER-SEM improved significantly after calibration, it represents only one data point. While the remaining platforms, having between five and seven data points, showed that the only statistic to generally improve (except for mil-ground C/C which had a lower PRED after calibration) was PRED (.25), yet even this improvement still does not meet Conte's guidelines. The rest of the statisics, when compared to the uncalibrated validation data set, actually got worse.

Having seen that SEER-SEM is calibratable, it is important to see if any bias existed in the estimates. This was accomplished by using Statistix 4.0. The validation data set was run through SEER-SEM (without selecting the Class knowledge base which would have included the calibration effort adjustment factor) and then compared to the actuals using the Wilcoxon signed rank test. The validation data set was then run through the model again but this time included the Class knowledge base and then compared to the actuals. Table 4.5 shows the results of this test. Military Specification Avionics is not included because it only had one validation data point.

Table 4.5 Wilcoxon Signed Rank Test Results

	Military Ground C/C	**Military Ground Signal Processing**	**Military Mobile**
Validation Data no factor	.2968	.8124	.875
Validation Data w/factor	.6874	.4688	1.125

In all platform applications tested, the model is biased. For example in military mobile we can only say with 12.5% confidence (1-.875) that the validation data for this platform is not biased. To determine if the estimates are either biased high or low the means of the differences between the no factor data set compared to the actuals and the factor data set compared to the actuals were calculated. Table 4.6 depicts the results of this effort. Again Military Specification Avionics is not included because of only having one validation point.

Table 4.6 Mean Difference of Validation Data Set
and
Validation Data Set (no factor) to Actuals

	Military Ground C/C	**Military Ground Signal Processing**	**Military Mobile**
Validation Data no factor	38.34	-47.60	36.40
Validation Data w/factor	17.05	-205.75	-11.58

The above data now shows the direction of bias contained in the validation data set when run through SEER-SEM. The results do not indicate constant bias across platforms. The military ground command/control platform indicates positive bias, SEER's estimates are low. While the military ground signal processing platform stratification shows negative bias, high estimates. Military mobile on the other hand indicates positive bias before calibration and negative bias after the calibration effort adjustment factor is applied to the data.

Summary

This chapter presented the data results and analysis of this research effort. The usefulness of the SWDB was addressed, the SEER-SEM was calibrated and SEER-SEM was validated against selected data points. The results of this effort indicate that the model is calibrateable, but the accuracy of the model was not in line with expectations. Chapter V presents conclusions about this research effort and suggests recommendations for further study.

V. Conclusions and Recommendations

Introduction

This chapter addresses the conclusions and recommendations of this research effort. The conclusions summarize the findings of this effort while the recommendations offer some ideas on further research possibilities that should be accomplished in this area.

Conclusions

Based upon the SEER-SEM User's Manual and an Air Force sponsored study, estimating accuracy of the SEER-SEM in this particular effort was significantly lower than anticipated. The SEER-SEM User's Manual claimed that model accuracy is normally within 10% of actuals even without calibration (SEER-SEM User's Manual, 1994: 11-7). Also, a Management Consulting & Research, Inc. report, sponsored by SMC, indicated that the overall ability of the SEER-SEM to estimate the effort required for then SSD programs was validated to within 5.86% (Apgar, 1991: III-1). This study could not replicate this degree of model accuracy. In fact, this study showed that model accuracy was rather limited and only marginally improved after model calibration. Again, the improvement came primariliy from the general improvement in the PRED (.25) statistic. The rest of the statistics deteriorate except for the Mil-Spec Avionics platform which only had one validation data point. The accuracy of SEER's estimates for the calibrated validation data set range from 29.6% to 161% of actual effort (See table 4.4 RRMS values). However, the limited accuracy displayed in this effort may not be because of any inherent defects in the model itself but from the diversification of the data used.

The data points selected for use in this effort were provided to SMC from multiple and most likely diverse organizations. The data sets used in this effort included Air Force, Army, and Navy records. Each Service has their own particular way of doing things and each has a unique mission. When varied data is grouped together for a model calibration it becomes highly unlikely that the results will be of use to any one organization. Too many factors are ignored when data sets represent such diversity. The calibration of an effort adjustment factor becomes only marginally beneficial when that factor is then applied across the spectrum of data points within each platform stratification..

Besides the Service affiliation of the data sets, the diversification of the organizations within each service creates problems with the continuity of the data. For example, many of the Air Force data sets had widely differing software project attributes. Not only are the people different from one Air Force organization to another but also the complexity of the software project and the project's intended mission differ. These factors limit the accuracy of effort calibration across organizational boundaries. This conclusion is supported by Bailey and Basili who found

> that due to the great variability of factors influencing software projects in different organizations, no cost models are truly transportable between environments; ultimately there can be no useful generic model. That does not mean that organizations should abandon software prediction but rather that each should examine its own environment carefully and produce or calibrate a model suited to its own requirements. (Bailey and Basili, 1981: 107-116)

Another factor that contributed to the limited accuracy of SEER-SEM involved the platforms themselves. During the calibration effort, SMC corrected the placement of many unmanned space records into a platform called ground in support of space. SEER does not include such a distinction. These data points ended up in the military ground platform stratifications. This would likely impact the accuracy of the estimates SEER provided because to this researcher, military ground is not as complex or as difficult a task to accomplish as a ground in support of space application.

The level of detail that a particular record provided also impacted the accuracy of the estimates of SEER-SEM. Such key factors as acquisition development method and development standard were often times not provided in the data sets used. Some records contained most of the available SWDB data fields while others contained little beyond platform, software application, size and effort. In situations where additional data fields were recorded, there could be significant differences in the final estimate that SEER-SEM provided. For example, a larger project in SLOC terms may, if it included the additional data fields, result in a smaller effort estimate than would a smaller project without the additional data. Therefore, the lack of critical data severely impacts the calculated calibration effort adjustment factors for each platform.

With regards to the model itself, SEER-SEM is a fairly easy model to use. It's user manual does a good job of describing the model's various functions and the steps necessary for estimate creation. However, the chapter on calibration was confusing to this researcher. Once given a tutorial from Galorath Associates, the actual calibration steps are rather straightforward and simple. This model is calibrateable and, based on this effort, responds to calibration. A calibration effort is worthwhile if detailed historical data is available and the data can be stratified by organization.

Recommendations

As a result of this research effort, it became apparent to this researcher that, although the model is not particularly accurate for this data or perhaps the way in which this data was used, the model does at least provide a basis for comparison to industry averages.

A significant effort needs to be done in the area of collecting useable data for calibration efforts. The SWDB provides basically good top-down information for calibration efforts, but what is needed is more detailed information. Is there some way of going beyond the information that is contained in the SWDB? Is there any way to get missing inputs into the existing records? Are there any requirements for providers of the data to complete the available fields before the information is added to the database? Can an additional field be included that identifies which organization is actually performing the development effort? It became obvious that many software developments were performed by the same offices. Adding a office field would improve the stratification process. This could be done without compromising the proprietary nature of the information and would likely improve the calibration of not only SEER but other software cost estimating models.

Another area of potential research effort is the creation, calibration and validation of a new platform selection within SEER-SEM, called ground in support of space. In this effort many of the data points originally used for unmanned space calibration/validation where later found to be ground in support of space applications. The SWDB already makes this distinction in its formatting of platform data; however, SEER-SEM does not.

In this research effort these unmanned space ground points were either included in military ground command and control or military ground signal processing. Neither selection is ideal because of the inherent complexities of unmanned space ground support systems over typical ground applications. This surely impacted the accuracy of SEER in these particular software environments.

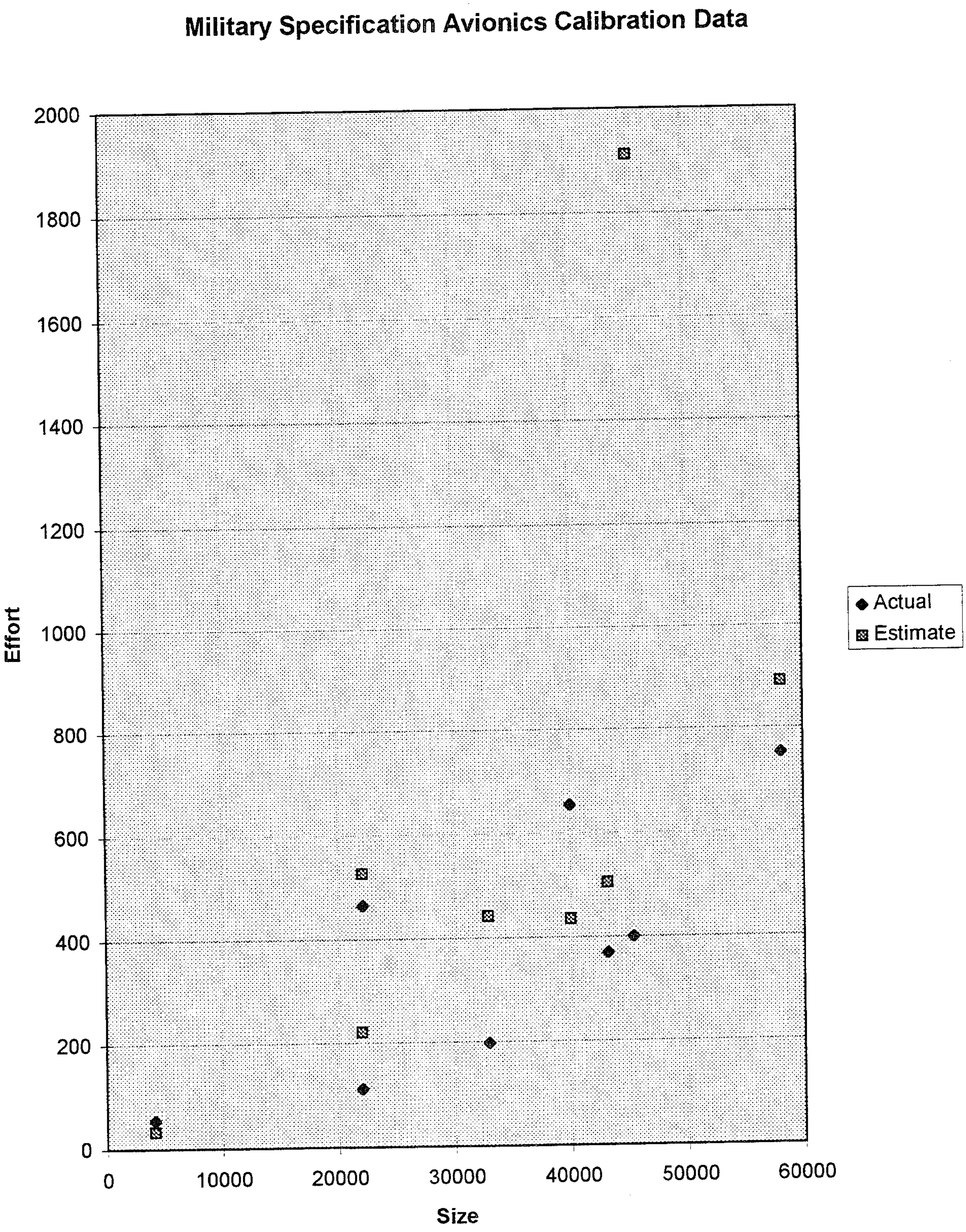
Software development size is yet another area where further research may yield important results in the application of SEER-SEM. Referencing the graphs in Appendix A, SEER-SEM's estimates of the actual effort are generally better for the smaller efforts than for the larger ones. Perhaps a stratification using both platform and size would give better results. Should a separate knowledge base be created in SEER-SEM that would split platforms into small and large software development efforts? The definitions of both small and large and even perhaps a medium category would have to be defined in a subsequent research effort.

Provided the historical data can be obtained, this effort should be repeated using both the calibration effort adjustment factor and the calibration schedule adjustment factor. Because almost all of the data points had incomplete schedule information, the \schedule adjustment factor was not used in this effort but if used may greatly improve the results of subsequent SEER-SEM calibration efforts.

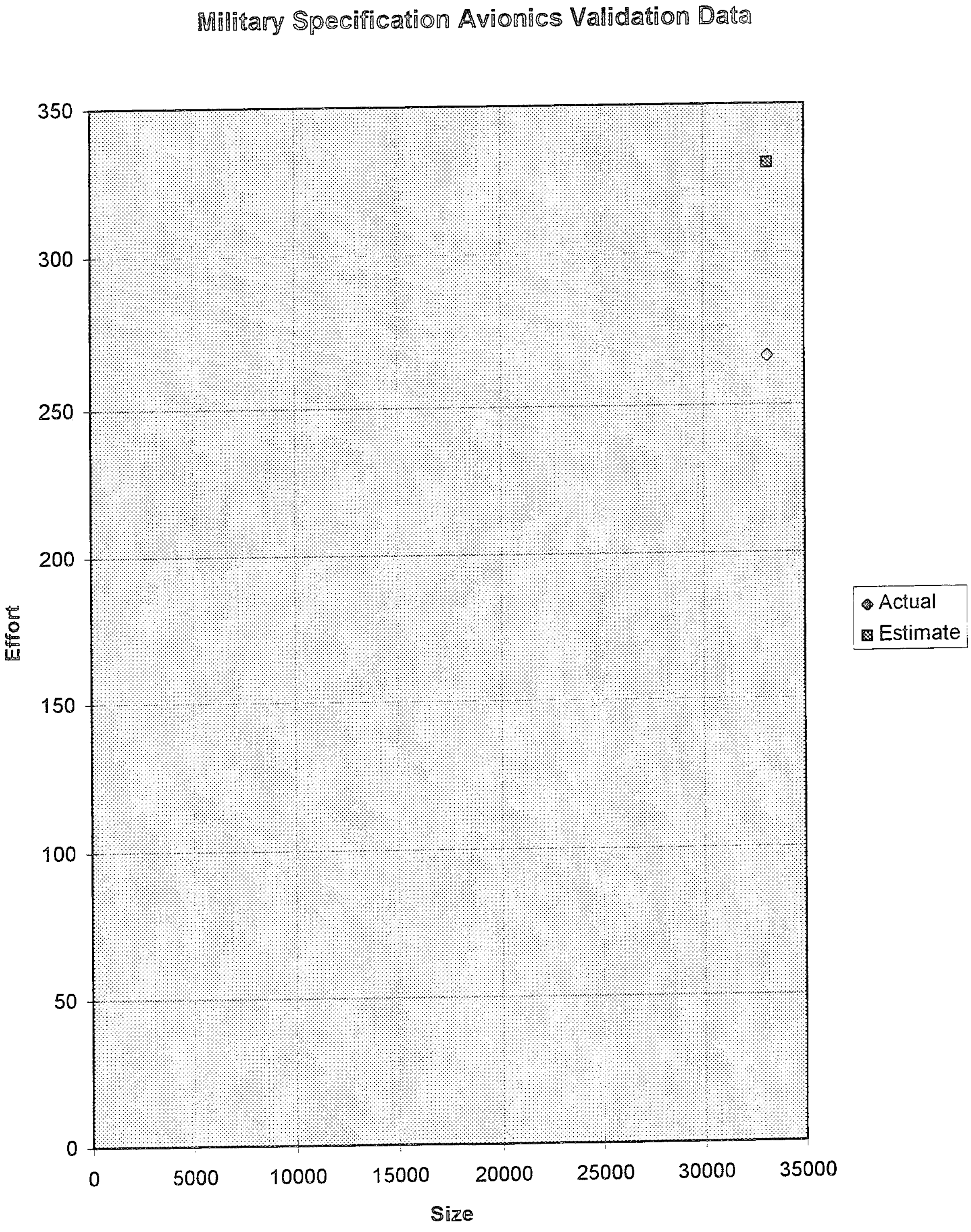
Summary

This chapter summarized the results of this effort. Conclusions as to the accuracy obtained were discussed and recommendations for additional study were provided.

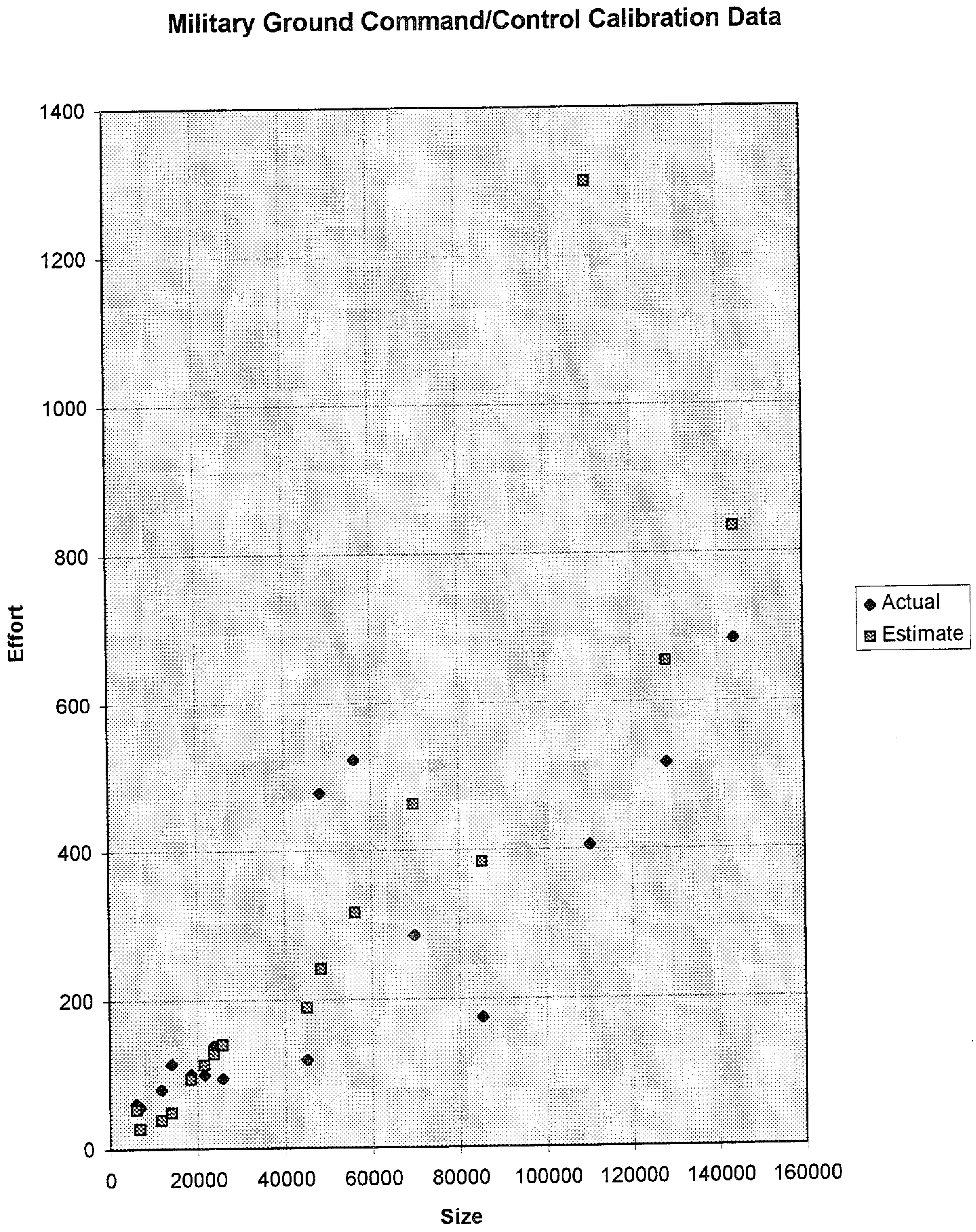
Appendix A. Platform Size to Effort Charts



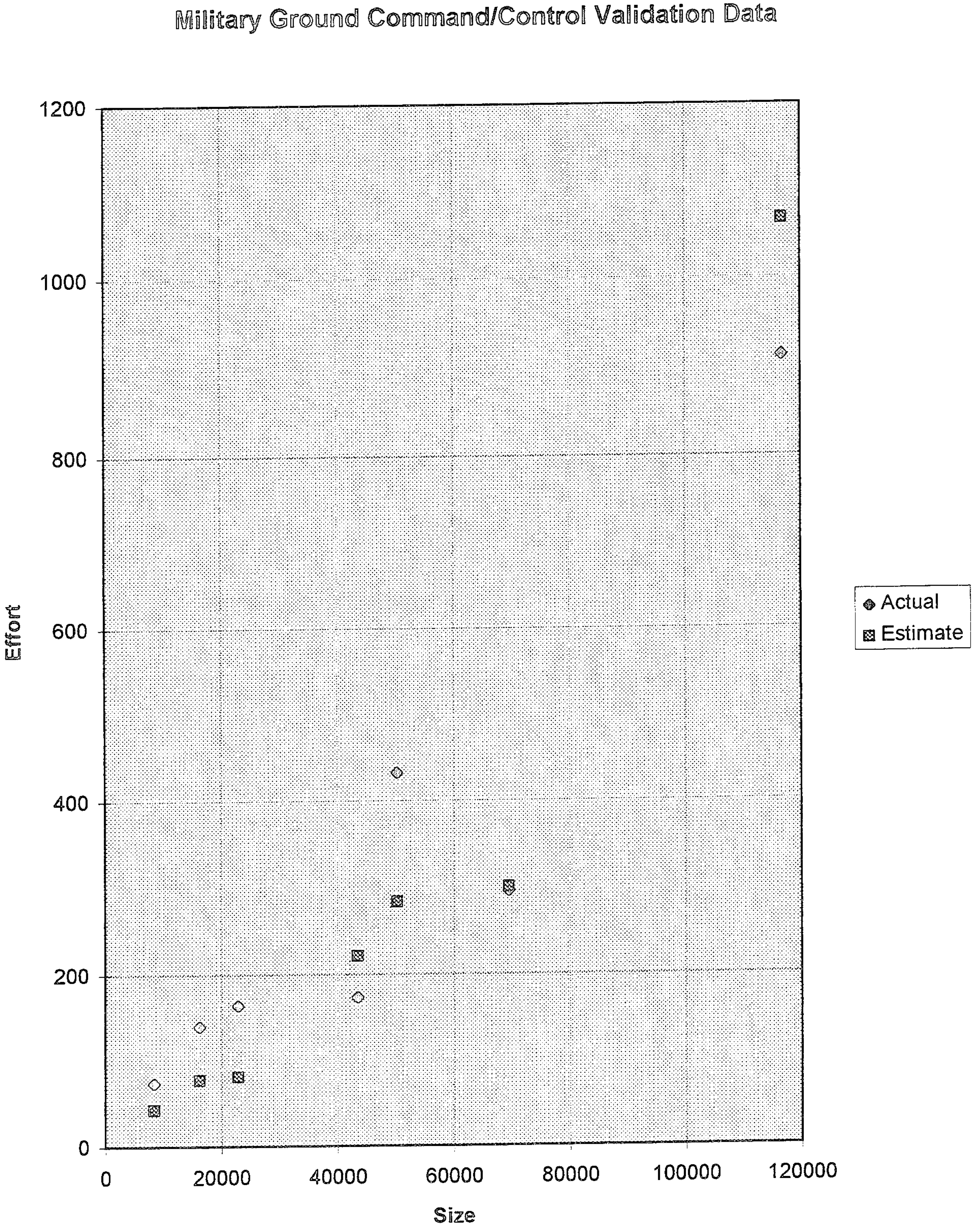
Appendix A. Platform Size to Effort Charts



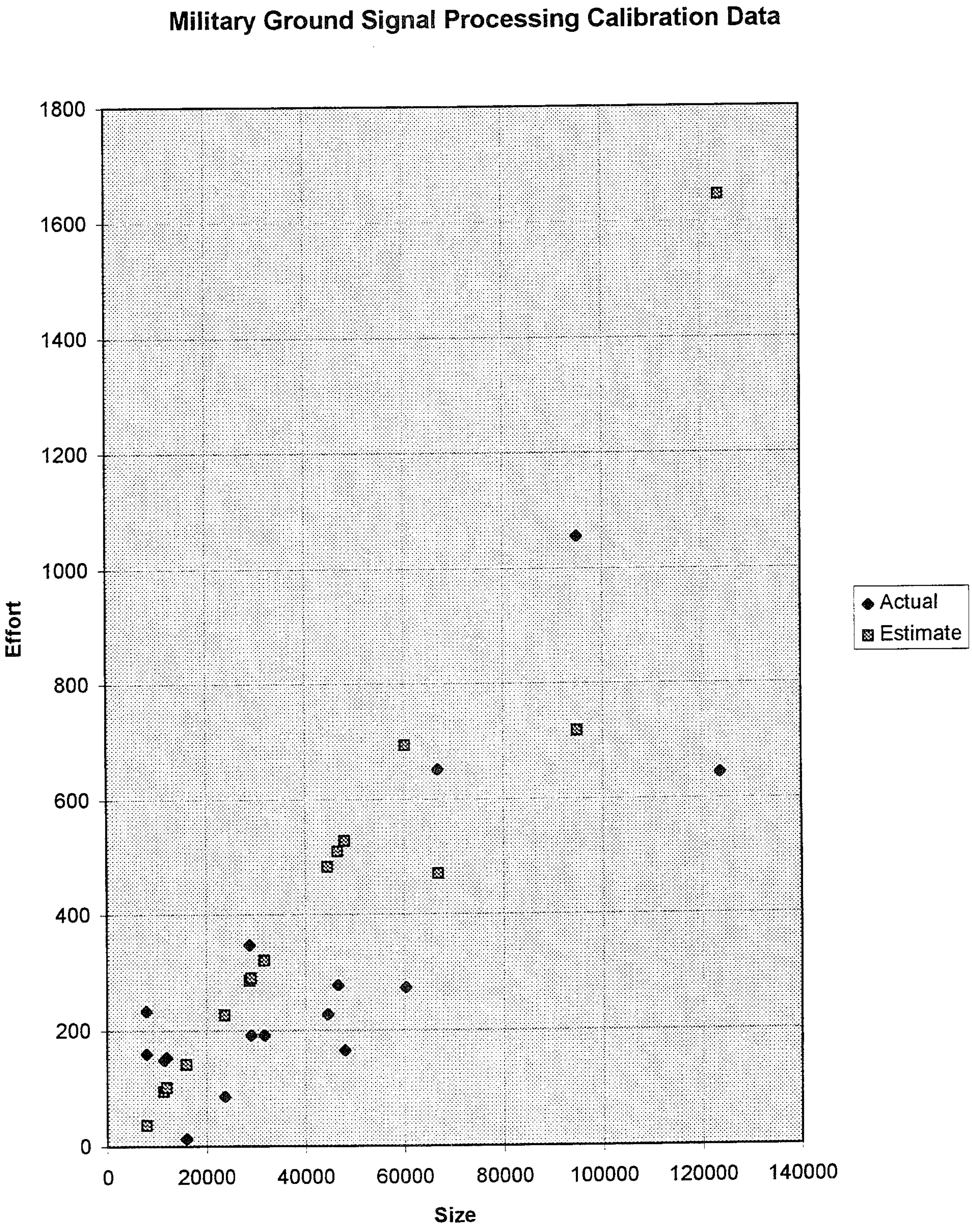
Appendix A. Platform Size to Effort Charts



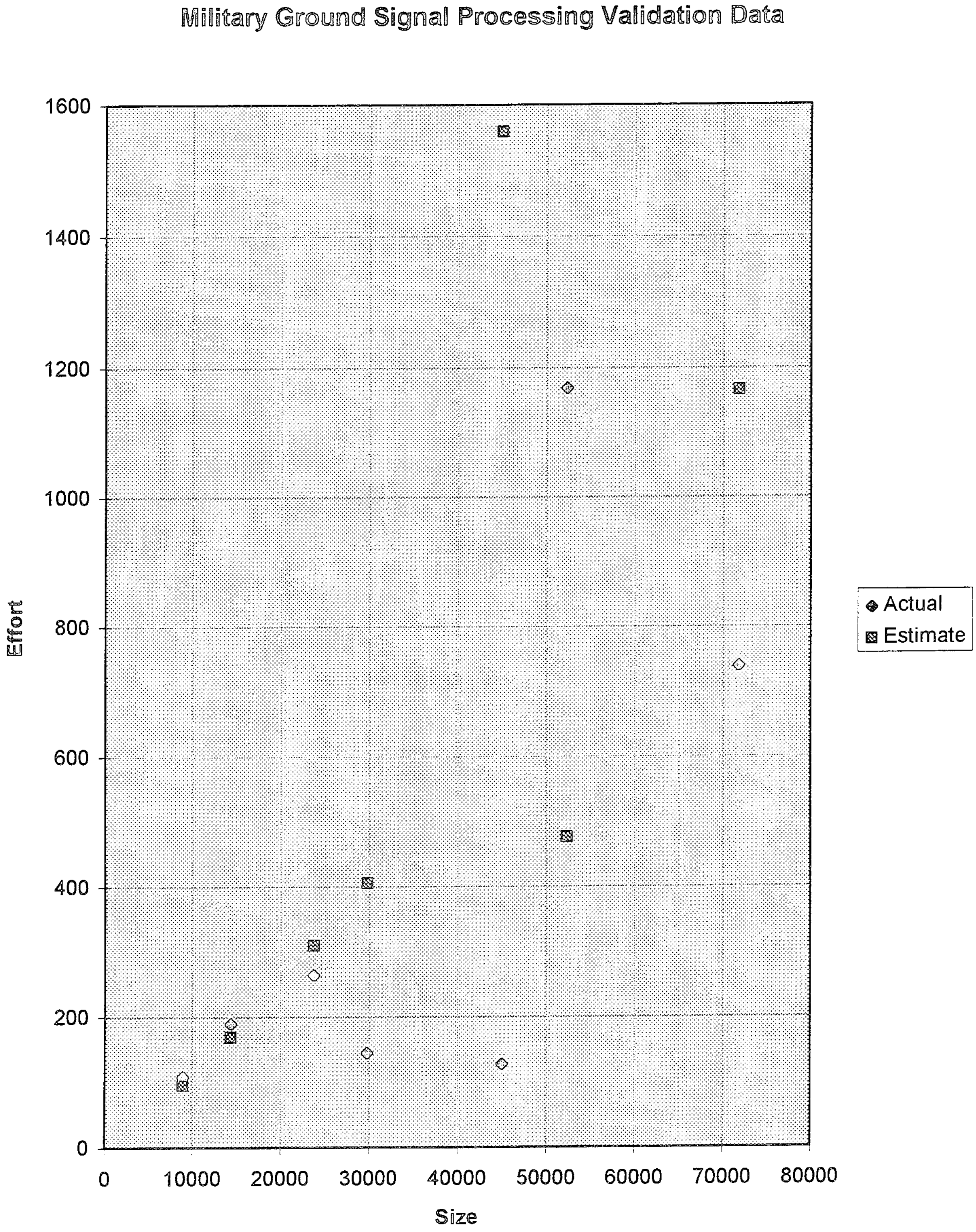
Appendix A. Platform Size to Effort Charts



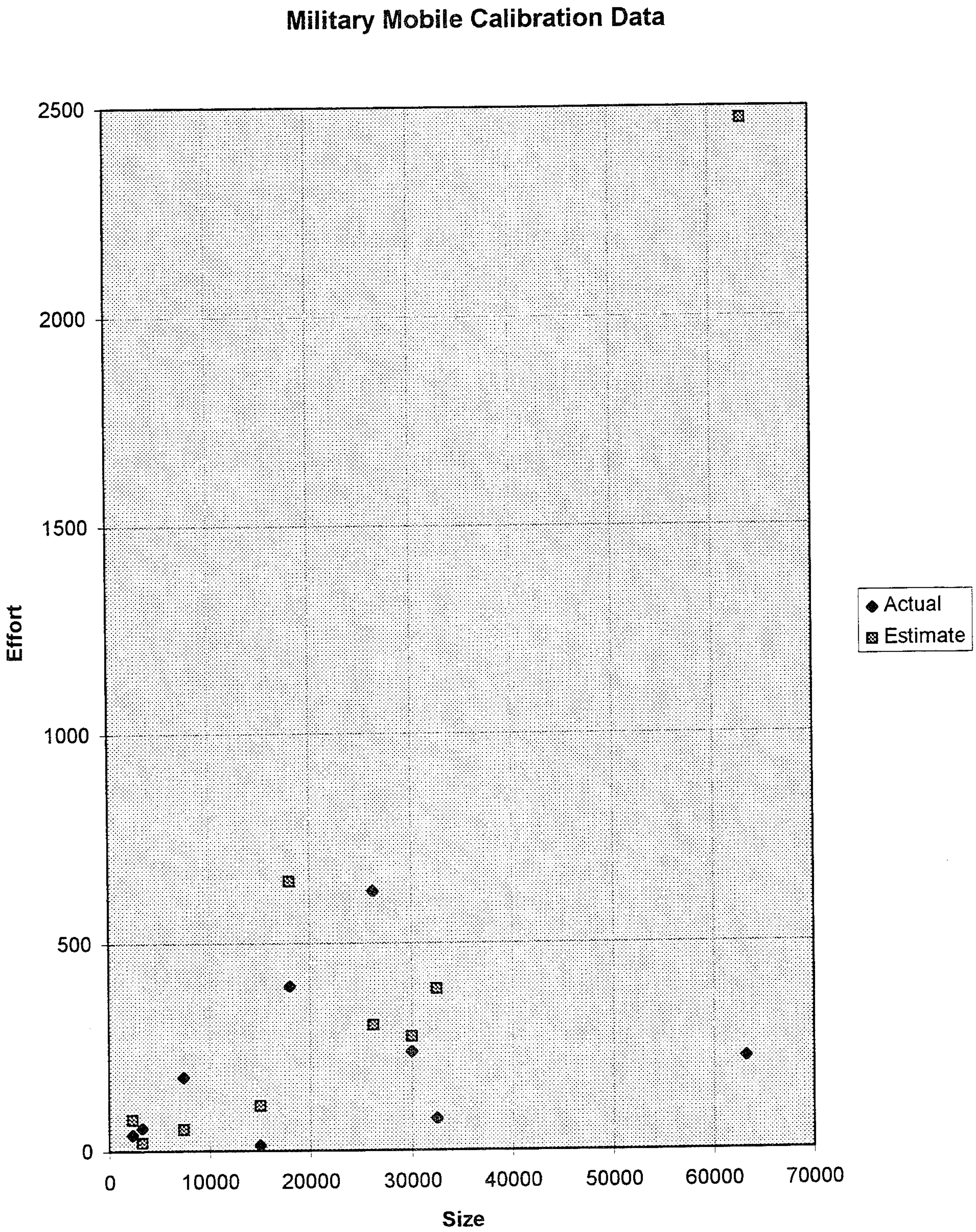
Appendix A. Platform Size to Effort Charts



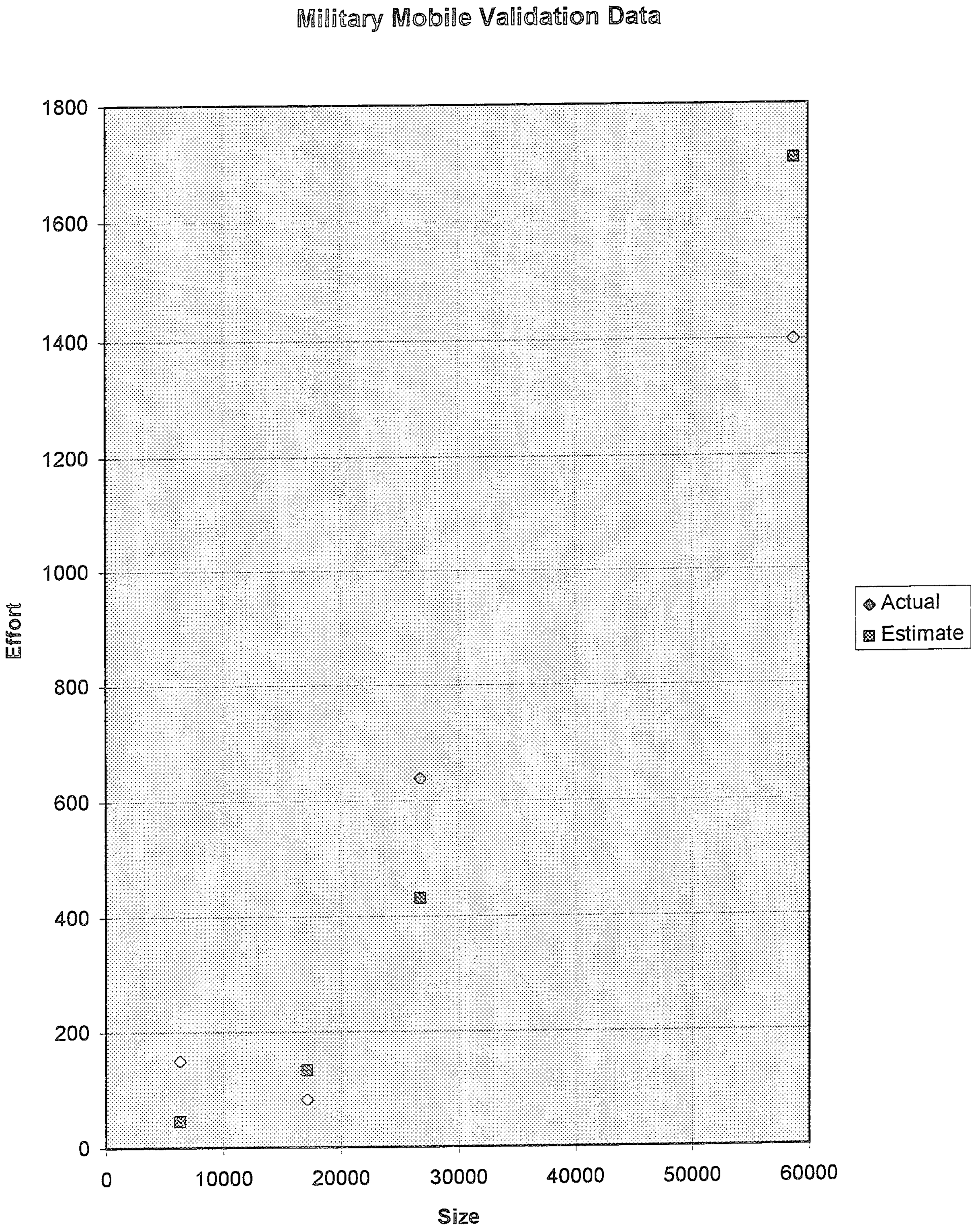
Appendix A. Platform Size to Effort Charts



Appendix A. Platform Size to Effort Charts



Appendix A. Platform Size to Effort Charts



Appendix B: Calibration/Validation Statistics Data

Military Specification Avionics

		Calibration							%		
Record	Size	Actual	Estimate	Delta	MRE	MMRE	RMS	RRMS	Change	Counter	
67	4144	54	32.63	21.37	0.396				0.3957	0	
12	22027	112	220.32	-108.32	0.967				-0.9671	0	
14	22148	464	525.39	-61.39	0.132				-0.1323	1	
11	32878	198	442.17	-244.17	1.233				-1.2332	0	
346	40000	654	435.04	218.96	0.335				0.3348	0	
10	43207	370	503.79	-133.79	0.362				-0.3616	0	
302	45353	400	1911.42	-1511.4	3.779				-3.7786	0	
13	58153	752	889.9	-137.9	0.183				-0.1834	1	
		3004		2444828	7.387	0.9233	552.8	1.472		25.00%	

		Comparison							%		
Record	Size	Actual	Estimate	Delta	MRE	MMRE	RMS	RRMS	Change	Counter	
*2512	33158	266	355.68	-89.726	0.337	0.459	89.73	0.337	-0.3374	0	
				8050.76						0.00%	
											Wilcoxon
		Validation							%		Compared
Record	Size	Actual	Estimate	Delta	MRE	MMRE	RMS	RRMS	Change	Counter	Data
*2512	33158	266	329.87	-63.916	0.24				-0.2403	1	388.08
		266		4085.26	0.24	0.2403	63.92	0.24		100.00%	

Military Ground Command/Control

		Calibration							%		
Record	Size	Actual	Estimate	Delta	MRE	MMRE	RMS	RRMS	Change	Counter	
38	6000	61	52.34	8.66	0.142				0.142	1	
83	6800	57	26.98	30.02	0.527				0.5267	0	
74	11700	80	38.81	41.19	0.515				0.5149	0	
76	14000	115	48.33	66.67	0.58				0.5797	0	
145	18560	101	94.33	6.67	0.066				0.066	1	
150	21681	100	113.67	-13.67	0.137				-0.1367	1	
124	23881	139	127.65	11.35	0.082				0.0817	1	
120	25842	95	140.32	-45.32	0.477				-0.4771	0	
7	45057	120	189.74	-69.74	0.581				-0.5812	0	
78	48300	478	241.32	236.68	0.495				0.4951	0	
77	56200	523	316.18	206.82	0.395				0.3954	0	
152	69772	286	462.13	-176.13	0.616				-0.6158	0	
*2517	85382	175.8	384.01	-208.22	1.184				-1.1845	0	
*2501	110400	405.2	1299.96	-894.78	2.208				-2.2083	0	
9	128200	517	654.26	-137.26	0.265				-0.2655	0	
50	144000	684	835.13	-151.13	0.221				-0.221	1	
		3937		1029869	8.492	0.5307	253.7	1.031		31.25%	

* Indicates that this record reported the total number of hours/person month differently than our assumed 152 hours/person month and was normalized accordingly.

Appendix B: Calibration/Validation Statistics Data

		Comparison							%		
Record	Size	Actual	Estimate	Delta	MRE	MMRE	RMS	RRMS	Change	Counter	
155	8398	74	36.42	37.58	0.508				0.5079	0	
82	16300	140	73.68	66.32	0.474				0.4737	0	
81	22900	164	76.91	87.09	0.531				0.531	0	
*2510	43437	173.1	206.32	-33.19	0.192				-0.1917	1	
79	50300	432	267.98	164.02	0.38				0.3797	0	
80	69450	296	282.96	13.04	0.044				0.0441	1	
75	116800	912	978.47	-66.47	0.073				-0.0729	1	
		2191		45987.7	2.201	0.3144	81.05	0.259		42.86%	
											Wilcoxon
		Validation							%		Compared
Record	Size	Actual	Estimate	Delta	MRE	MMRE	RMS	RRMS	Change	Counter	Data
155	8398	74	42.61	31.39	0.424				0.4242	0	36.42
82	16300	140	77.79	62.21	0.444				0.4444	0	66.49
81	22900	164	81.24	82.76	0.505				0.5046	0	69.44
*2510	43437	173.1	220.9	-47.768	0.276				-0.2759	0	188.80
79	50300	432	283.13	148.87	0.345				0.3446	0	241.99
80	69450	296	299.86	-3.86	0.013				-0.013	1	256.29
75	116800	912	1066.26	-154.26	0.169				-0.1691	1	911.33
		2191		59959.8	2.176	0.3108	92.55	0.296		28.57%	
Military Ground Signal Processing											
		Calibration							%		
Record	Size	Actual	Estimate	Delta	MRE	MMRE	RMS	RRMS	Change	Counter	
99	8000	234	36.86	197.14	0.842				0.8425	0	
107	8000	160	36.86	123.14	0.77				0.7696	0	
153	11534	149	95.49	53.51	0.359				0.3591	0	
136	12121	154	101.35	52.65	0.342				0.3419	0	
127	16016	13	141.59	-128.59	9.892				-9.8915	0	
143	23703	86	226.64	-140.64	1.635				-1.6353	0	
142	28782	348	286.1	61.9	0.178				0.1779	1	
131	29147	192	290.46	-98.46	0.513				-0.5128	0	
147	31720	192	321.49	-129.49	0.674				-0.6744	0	
134	44527	228	482.96	-254.96	1.118				-1.1182	0	
132	46595	278	510	-232	0.835				-0.8345	0	
126	47965	165	528.05	-363.05	2.2				-2.2003	0	
137	60233	274	694.01	-420.01	1.533				-1.5329	0	
117	66843	652	470.83	181.17	0.278				0.2779	0	
90	95000	1055	717.9	337.1	0.32				0.3195	0	
133	123710	645	1646.09	-1001.1	1.552				-1.5521	0	
		4825		1701954	23.04	1.44	326.1	1.082		6.25%	

* Indicates that this record reported the total number of hours/person month differently than our assumed 152 hours/person month and was normalized accordingly.

Appendix B: Calibration/Validation Statistics Data

		Comparison							%		
Record	Size	Actual	Estimate	Delta	MRE	MMRE	RMS	RRMS	Change	Counter	
154	8965	109	70.57	38.4265	0.353				0.3525	0	
138	14389	190	124.51	65.4926	0.345				0.3447	0	
135	23787	264	227.60	36.3971	0.138				0.1379	1	
144	29802	145	298.31	-153.31	1.057				-1.0573	0	
54	45035	127	1146.05	-1019.1	8.024				-8.024	0	
91	52275	1169	350.55	818.449	0.7				0.7001	0	
130	71851	738	857.60	-119.6	0.162				-0.1621	1	
		2742		1753223	10.78	1.5398	500.5	1.278		28.57%	
											Wilcoxon
		Validation							%		Compared
Record	Size	Actual	Estimate	Delta	MRE	MMRE	RMS	RRMS	Change	Counter	Data
154	8965	109	95.98	13.02	0.119				0.1194	1	70.57
138	14389	190	169.33	20.67	0.109				0.1088	1	124.51
135	23787	264	309.54	-45.54	0.173				-0.1725	1	227.60
144	29802	145	405.7	-260.7	1.798				-1.7979	0	298.31
54	45035	127	1558.63	-1431.6	11.27				-11.273	0	1146.05
91	52275	1169	476.75	692.25	0.592				0.5922	0	350.55
130	71851	738	1166.34	-428.34	0.58				-0.5804	0	857.60
		2742		2782885	14.64	2.092	630.5	1.61		42.86%	
					Military Mobile						
		Calibration							%		
Record	Size	Actual	Estimate	Delta	MRE	MMRE	RMS	RRMS	Change	Counter	
347	2311	39	75.73	-36.73	0.942				-0.9418	0	
349	3268	56	19.29	36.71	0.656				0.6555	0	
*2505	7448	177.6	52.08	125.552	0.707				0.7068	0	
2515	15025	13	108.87	-95.87	7.375				-7.3746	0	
348	18052	396	648.89	-252.89	0.639				-0.6386	0	
*2504	26239	624.7	301.26	323.411	0.518				0.5177	0	
303	30000	237	272.92	-35.92	0.152				-0.1516	1	
*2503	32464	76.97	388.95	-311.98	4.053				-4.053	0	
2456	63254	221	2470.89	-2249.9	10.18				-10.18	0	
		1841		5356824	25.22	2.8022	771.5	3.771		11.11%	
		Comparison							%		
Record	Size	Actual	Estimate	Delta	MRE	MMRE	RMS	RRMS	Change	Counter	
*2506	6317	150	42.18	107.82	0.719				0.7188	0	
34	17134	83	111.31	-28.31	0.341				-0.3411	0	
*2507	26814	638.5	398.11	240.377	0.376				0.3765	0	
*2508	58789	1399	1574.02	-174.68	0.125				-0.1248	1	
		2271		100720	1.561	0.3903	158.7	0.28		25.00%	

* Indicates that this record reported the total number of hours/person month differently than our assumed 152 hours/person month and was normalized accordingly.

Appendix B: Calibration/Validation Statistics Data

											Wilcoxon
		Validation							%		Compared
Record	Size	Actual	Estimate	Delta	MRE	MMRE	RMS	RRMS	Change	Counter	Data
*2506	6317	150	45.7	104.3	0.695				0.6953	0	38.08
34	17134	83	133.58	-50.58	0.609				-0.6094	0	111.32
*2507	26814	638.5	431.64	206.847	0.324				0.324	0	359.70
*2508	58789	1399	1706.22	-306.88	0.219				-0.2193	1	1421.85
		2271		150396	1.848	0.462	193.9	0.342		25.00%	

* Indicates that this record reported the total number of hours/person month differently than our assumed 152 hours/person month and was normalized accordingly.

Appendix C: Calibration Effort Adjustment Factors

Military Specification Avionics - Calibration

Record	Size	Actuals	Estimate	Adjustment factor
67	4144	54	32.63	1.65
12	22027	112	220.32	0.51
14	22148	464	525.39	0.88
11	32878	198	442.17	0.45
346	40000	654	435.04	1.50
10	43207	370	503.79	0.73
302	45353	400	1911.42	0.21
13	58153	752	889.9	0.85
			Mean=	0.85

Military Ground Command/Control - Calibration

Record	Size	Actuals	Estimate	Adjustment factor
38	6000	61	52.34	1.17
83	6800	57	26.98	2.11
74	11700	80	38.81	2.06
76	14000	115	48.33	2.38
145	18560	101	94.33	1.07
150	21681	100	113.67	0.88
124	23881	139	127.65	1.09
120	25842	95	140.32	0.68
7	45057	120	189.74	0.63
78	48300	478	241.32	1.98
77	56200	523	316.18	1.65
152	69772	286	462.13	0.62
*2517	85382	175.78947	384.01	0.46
*2501	110400	405.18421	1299.96	0.31
9	128200	517	654.26	0.79
50	144000	684	835.13	0.82
			Mean =	1.17

* Indicates that this record reported the total number of hours/person month differently than our assumed 152 hours/person month and was normalized accordingly.

Appendix C: Calibration Effort Adjustment Factors

Military Ground Signal Processing - Calibration

Record	Size	Actuals	Estimate	Adjustment factor
99	8000	234	36.86	6.35
107	8000	160	36.86	4.34
153	11534	149	95.49	1.56
136	12121	154	101.35	1.52
127	16016	13	141.59	0.09
143	23703	86	226.64	0.38
142	28782	348	286.1	1.22
131	29147	192	290.46	0.66
147	31720	192	321.49	0.60
134	44527	228	482.96	0.47
132	46595	278	510	0.55
126	47965	165	528.05	0.31
137	60233	274	694.01	0.39
117	66843	652	470.83	1.38
90	95000	1055	717.9	1.47
133	123710	645	1646.09	0.39
			Mean =	1.36

Military Mobile - Calibration

Record	Size	Actuals	Estimate	Adjustment factor
347	2311	39	75.73	0.51
349	3268	56	19.29	2.90
*2505	7448	177.63158	52.08	3.41
2515	15025	13	108.87	0.12
348	18052	396	648.89	0.61
*2504	26239	624.67105	301.26	2.07
303	30000	237	272.92	0.87
*2503	32464	76.973684	388.95	0.20
2456	63254	221	2470.89	0.09
			Mean =	1.20

* Indicates that this record reported the total number of hours/person month differently than our assumed 152 hours/person month and was normalized accordingly.

Bibliography

Apgar, Henry et al. Application Oriented Software Data Collection Software Model Calibration Report. Contract F33657-87-D-0107. Oxnard CA: Management Consulting & Research, Inc., 15 March 1991 (TR-9007/49-1).

Bailey, J. W. and Basili, V. R. "A Meta-Model for Software Development Resource Expenditure," Proceedings of the 5th International Conference on Software Engineering IEEE/ACM/NBS. 107-116. March 1981.

Boehm, Barry W. "Improving Software Productivity," Computer, 43-57 (September 1987.

Boehm, Barry W. "Software Engineering Economics," IEEE Transactions on Software Engineering, 10: 4-21 (January 1984).

Christensen, David S. and Daniel V. Ferens. "Using Earned Value for Performance Measurement on Software Development Projects," Dayton OH. Air Force Institute of Technology, 15 April 1993.

Coggins, George A. and Roy C. Russell. Software Cost Estimating Models: A Comparative Study of What the Models Estimate. MS thesis, AFIT/GCA/LAS/93S-4. School of Systems and Logistics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1993 (AD-A275989).

Conte, S.D., H. E. Dunsmore, and V. Y Shen, Software Engineering Metrics and Models. Menlo Park CA: Benjamin/Cummings Publishing Company, Inc., 1986.

Dreger, J. B., Function Point Analysis. Englewood Cliffs NJ: Prentice Hall, 1989.

Ferens, Daniel V. Class handout, Cost 677, Quantitative Management of Software. School of Systems and Logistics, Air Force Institute of Technology, Wright-Patterson AFB OH, November 1994.

Ferens, Daniel V. and Robert B. Gurner, "An Evaluation of Three Function Point Models for Estimation of Software Effort," Course Text, Cost 677, Quantitative Management of Software. School of Systems and Logistics, Air Force Institute of Technology, Wright-Patterson AFB OH, November 1994.

Fulton, Richard and Sherry Stukes, Space and Missile Systems Center Software Database User's Manual, Version 1.0. Management Consulting & Research, Inc., Oxnard CA, September 1993.

Gibbs, W. Wyatt. "Software's Chronic Crisis," Scientific American: 86-95, September 1994.

Glass, Robert L. "The Software Crisis...Not?" Computer, 104, April 1994.

Humphrey, Watts S. Managing the Software Process. Reading MA: Addison-Wesley Publishing Company, 1989.

Jones, Capers, "Software Metrics: Good, Bad, and Missing," Computer, 27: 98-100, September 1994.

Kemerer, Chris F. "An Empirical Validation of Software Cost Estimation Models," Communications of the ACM, 30: 416-429, May 1987.

Kitfield, James, "Is Software DoD's 'Achilles' Heel," Military Forum, 28-34 (July 1989).

Latamore, G. Berton and Joseph Maglitta. "Riding the Software Pricing Skyrocket," Computerworld: 99-100, 2 November 1992.

Londeix, Bernard Cost Estimation for Software Development. Wokingham, England: Addison-Wesley Publishing Company, 1987.

Marsh, Alton, "Pentagon Up Against a Software Wall," Government Executive, 62-63 (May 1990).

McRitchie, Karen. SEER Technologies Division, Galorath Associates Incorporated, Los Angeles CA. Personal interview. 12-14 July 1995.

Mendenhall, William, Dennis D. Wackerly, and Richard L Scheaffer. Mathematical Statistics with Applications (Fourth Edition). Belmont CA: Duxbury Press, 1990.

Novak-Ley, Gina. Model Calibration Task Monitor, Space and Missile Center, Los Angeles CA. Personnel interview. 12 October 1994.

Ourada, Gerald, L. Software Cost Estimating Models: A Calibration, Validation and Comparison. MS thesis, AFIT/GSS/LSY/91D-11. School of Systems and Logistics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, December 1991 (AD-A246677).

SEER-SEM User's Manual, Version 4.0, Galorath Associates Inc., Los Angeles CA, September 1994.

Statistix User's Manual, Version 4.0, Analytical Software, Tallahassee FL, 1992.

Stukes, Sherry and Henry Apgar. Air Force Cost Analysis Agency Software Model Content Study. Contract F19628-88-D-0002, Task 51. Oxnard CA: Management Consulting & Research, Inc., 18 April 1994 (TR-9359/51-8).

Stukes, Sherry and Henry Apgar. Space and Missile Systems Center Software Development Database (Phase Five) Final Report. Contract F04701-90-D-0002, Task 025. Oxnard CA: Management Consulting & Research, Inc., 1 November 1994 (TR-9338/025-02).

Symons, Charles R. Software Sizing and Estimating Mk II (FPA) Function Point Analysis. Chichester, England: John Wiley & Sons Ltd, 1991.

Wellman, Frank, Software Costing. New York: Prentice Hall, 1992.

Glossary:

Capability Maturity Model	(CMM)
Coefficient of Determination	(R^2)
Computer Software Component	(CSC)
Computer Software Unit	(CSU)
Computer System Configuration Item	(CSCI)
Cost Model	(COSTMODL)
Cost Estimating Relationships	(CERs)
Department of Defense	(DoD)
International Function Point User's Group	(IFPUG)
Magnitude of Relative Error	(MRE)
Mean Magnitude of Relative Error	(MMRE)
Prediction Test Level	(PRED)
Programmed Review of Information for Costing and Evaluation - Software	(PRICE-S)
Relative Root Mean Square Error	(RRMS)
Revised Enhanced Version of Intermediate COCOMO	(REVIC)
Root Mean Square Error	(RMS)
Software Architecture, Sizing and Estimating Tool	(SASET)
Software Life Cycle Model	(SLIM)
Source Line of Code	(SLOC)
Space and Missile Systems Center	(SMC)

Space and Missile Systems Center Software Database	(SWDB)
Space System's Cost Analysis Group	(SSCAG)
System Evaluation and Estimation of Resources Software Estimation Model	(SEER-SEM)
System Program Office	(SPO)

Vita

Captain Kolin D. Rathmann was born on 3 December 1966 in Milwaukee, Wisconsin. He graduated from Troy High School in 1985, completed the United States Air Force Preparatory School in May 1986, and graduated from the United States Air Force Academy in May 1990 with a Bachelor of Science in Business Administration and Management. He received his commission on 30 May 1990 upon graduation from the United States Air Force Academy. His first assignment was at Patrick AFB as a cost analyst for the 45th Space Wing. His second assignment was also at Patrick AFB but assigned to the tenant organization, Air Force Technical Applications Center (AFTAC), as a program analyst for the Center's nuclear treaty monitoring mission. In May 1994 he entered the Graduate School of Logistics and Acquisition Management, Air Force Institute of Technology for the Graduate Cost Analysis program.

Permanent Address: 7404 N. Hillrose PL
Peoria, IL 61614

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
September: 1995

3. REPORT TYPE AND DATES COVERED
Master's Thesis

4. TITLE AND SUBTITLE
CALIBRATION OF THE SYSTEM EVALUATION AND ESTIMATION OF RESOURCES SOFTWARE ESTIMATION MODEL (SEER-SEM) FOR THE AIR FORCE SPACE AND MISSILE SYSTEMS CENTER (SMC)

5. FUNDING NUMBERS

6. AUTHOR(S)
Kolin D. Rathmann, Capt, USAF

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology,
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/GCA/LAS/95S-9

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

SMC/FMC
2430 El Segundo Blvd, Ste 2010
El Segundo, CA 90245-4687

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This study examined whether calibration of the SEER-SEM impacted the effort estimates generated by the model for software developments. A historical database was provided by the Space and Missile Systems Center, Los Angeles, and used as the model's input data. The data was stratified into four usable platforms, military ground command/control, military ground signal processing, military specification avionics, and military mobile. Each platform's data sets were split, the majority of points for calibration of the model, and the rest for model validation. The accuracy of SEER to this particular data set is limited, yet the model did respond to calibration. It is recommended that further calibration attempts be done within specific organizations. The diversity of the SWDB created too many factors for SEER to overcome.

14. SUBJECT TERMS
Software Cost Model; Software Cost Estimates;
Software Cost Analysis; Computers

15. NUMBER OF PAGES
71

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UL